AI Circular Financing and the Coming Shakeout: Vendor Funding, Round-Tripping, and Hyperscaler Power in the Data-Center Boom

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

January 7, 2026

A new kind of AI bubble story is forming, but it is not centered on dot-com-style websites or vaporware. It is built from physical infrastructure: data centers, power contracts, GPUs, and long-dated capacity commitments. In this environment, money can move in loops. Hyperscalers and chip suppliers invest in frontier AI companies, those companies then spend large portions of that capital back on the hyperscalers' cloud and the suppliers' hardware, and intermediaries package the resulting cash flows into credit structures that look stable until utilization disappoints. The result is not necessarily fraud. It is an incentive-consistent arrangement that can still become fragile when growth slows. This paper defines what "circular deals" are, separates benign platform strategy from balance-sheet optics, and maps the most plausible crash pathways: demand shortfall, utilization gaps, repricing of compute, collateral devaluation, and forced deleveraging in the private-credit layer. It also treats market-structure implications as first-class: if distribution and default surfaces matter more than model quality, a small number of incumbents may consolidate control even as the ecosystem appears to diversify. The goal is a disciplined framework for evaluating risk without resorting to hype.

Keywords: artificial intelligence, AI bubble, circular financing, round-tripping, vendor financing, hyperscalers, cloud computing, data centers, GPUs, Nvidia, Google, Microsoft, Amazon Web Services, OpenAI, CoreWeave, private credit, asset-backed securities, capex cycle, utilization risk, systemic risk, market structure, antitrust, platform power.

A collaboration with GPT-5.2-Thinking. 51 pages. CC4.0.

Outline

1. Introduction
1.1 The shift from software narrative to infrastructure reality
1.2 Why “circular deals” became a mainstream concern
1.3 What this paper does and does not claim

2. The AI Capital Stack in 2024–2026
2.1 Who pays for what: chips, cloud, data centers, power, networks
2.2 The role of long-dated commitments and capacity reservations
2.3 The emergence of the neocloud layer

3. Defining Circularity
3.1 Circular financing vs normal strategic partnerships
3.2 Round-tripping patterns: investor-supplier-customer loops
3.3 Where circularity shows up in contracts and cash flows
3.4 A taxonomy of deal types and their risk signatures

4. Why Rational Actors Choose the Loop
4.1 Scarcity economics: GPUs, power, permitting, and time-to-ship
4.2 Competitive urgency: spending as a defensive moat
4.3 Distribution advantages: default surfaces and bundled adoption
4.4 Narrative arbitrage: valuation support and “proof of demand” optics

5. Where Fragility Hides
5.1 Leverage and maturity mismatch in the financing chain
5.2 Utilization risk: take-or-pay, minimum commits, and ramp schedules
5.3 Collateral dynamics: what a GPU is worth in a downturn
5.4 Accounting and reporting pressure points that can amplify shocks

6. Crash Pathways
6.1 The slow crash: ROI disappointment and capex deceleration
6.2 The fast crash: credit tightening and forced deleveraging
6.3 Secondary-market cascades: price discovery for compute and hardware
6.4 Spillovers: real estate, power markets, and local finance exposure

7. Consolidation Thesis
7.1 How incumbents can squeeze the field without “winning the models”
7.2 Google’s potential choke points: search, ads, Chrome, Android, workspace
7.3 Cloud platform gravity: switching costs and integration bundling
7.4 Counterforces: multi-cloud strategies, open weights, and specialization

8. Stress Tests and Simple Metrics
8.1 Unit economics of inference and the price floor for compute
8.2 Utilization sensitivity: what breaks when ramps miss
8.3 Credit signals: covenants, spreads, refinancing walls, concentration
**8.4 Competition signals: distribution capture vs model commoditization

9. Governance and Policy Options
9.1 Disclosure: standardizing commitments, utilization, and counterparty risk
9.2 Market integrity: limiting misleading demand signaling
9.3 Competition: interoperability, bundling scrutiny, and procurement rules
9.4 Safety and resilience: auditability without freezing innovation

10. Conclusion
10.1 The central claim: circularity is incentive-consistent but can be brittle
10.2 What would falsify the crash narrative
10.3 What would confirm it early enough to matter
10.4 Practical next steps for analysts, operators, and policymakers

References
R1 Reporting and analysis on circular AI deals and round-tripping
R2 Cloud capex cycles, data-center economics, and power constraints
R3 Private credit, structured finance, and systemic risk transmission
R4 Competition and platform economics in search and cloud AI
R5 Accounting, disclosure, and contract-risk literature

Art

1. Introduction

The public story of AI has mostly been told as a software story: new models, new benchmarks, new products, new magic. But the binding constraint of the current wave is increasingly not an algorithmic trick. It is industrial capacity. Training and serving frontier systems requires dense compute, high-bandwidth networking, massive storage, specialized cooling, and reliable power at a scale that forces choices about geography, permitting, and long-term procurement. When the center of gravity shifts from code to concrete, the economics change. The risk changes. The competitive logic changes. And the "bubble" question stops being about consumer apps and starts being about balance sheets, credit markets, and the durability of long-dated infrastructure bets.

This paper argues that many of the most controversial features of the current AI boom are not mysterious. They are predictable responses to scarcity, speed, and platform power. The important question is not whether these arrangements are "real." The question is whether they are robust under disappointment: slower revenue, lower utilization, cheaper compute, tighter credit, or a strategic consolidation that leaves some participants holding stranded commitments.

1.1 The shift from software narrative to infrastructure reality

For roughly two decades, the tech sector's most celebrated successes looked "light": software scaled on the back of commodity servers, and distribution rode the internet's marginal cost curve downward. That template shaped how many people intuitively think about AI. If models are software, then the story should be about product-market fit, virality, and subscription revenue.

The current AI cycle breaks that intuition in three ways.

First, capability at the frontier is compute-intensive and time-sensitive. The relevant competition is not merely "who writes the best code," but "who can secure the most compute the fastest," because delays are strategic losses. This converts technical ambition into procurement urgency.

Second, the supply chain is constrained. Advanced accelerators, high-end networking, and suitable data-center capacity do not appear on demand. They require manufacturing throughput, long lead times, and specialized integration. Even when capital is available, physical bottlenecks can dominate.

Third, the cost structure is not dominated by marginal software distribution. Serving AI at scale turns inference into a recurring industrial expense with a measurable unit economics. It looks

less like “ship code once” and more like “operate a fleet continuously.” That fleet has real depreciation schedules, power contracts, cooling limits, and upgrade cycles.

As a result, AI spending increasingly expresses itself as capital expenditures and long-dated commitments: data centers, equipment leases, power purchase agreements, reserved capacity, and multi-year compute deals. When those commitments become large relative to near-term revenue, the risk posture starts to resemble earlier infrastructure booms: telecom buildouts, energy cycles, and real estate expansions. The key failure mode is not “the app goes away.” It is “the utilization assumptions were wrong,” which is a different kind of fragility.

This is the backdrop for the rise of an intermediary layer: specialist infrastructure firms that aggregate GPUs, data-center space, and financing, then sell compute to model builders and enterprises. This layer can make AI growth feasible, but it also introduces classic credit-cycle dynamics: leverage, refinancing needs, covenants, collateral valuation, and counterparty concentration.

1.2 Why “circular deals” became a mainstream concern

The term “circular deals” entered mainstream discussion because observers noticed a pattern that looks unusual when described plainly: the same small group of companies can appear as investors, suppliers, and customers in a loop.

The basic structure is simple.

A capital-rich platform or supplier funds an AI developer or infrastructure partner. The recipient then spends much of that money back with the funder in the form of cloud services, compute, GPUs, networking gear, or reserved capacity. Sometimes the “return flow” is explicit in the deal design, sometimes it emerges from market constraints: there are only so many places to buy the required compute at scale, so the money naturally cycles back to the dominant providers.

This became salient for several reasons.

One reason is optics. In a hot market, “demand” can be signaled by headline commitments and large procurement contracts. If those contracts are effectively financed by the supplier or a related party, observers worry that demand is being amplified by the financing structure rather than by independent end-customer pull. Even when everything is disclosed and legal, the psychological effect is the same: the market may misread the strength and breadth of demand.

A second reason is concentration. Circularity is easier when a small number of firms control scarce inputs: chips, cloud capacity, distribution surfaces, and enterprise channels. The loop is, in part, a symptom of oligopoly conditions. That naturally raises anxiety about consolidation: if the dominant platforms are simultaneously capital providers and distribution gatekeepers,

challengers may become structurally dependent before they become economically self-sustaining.

A third reason is credit risk. Circularity becomes more concerning when it is paired with leverage and structured finance. If the “return flow” funds debt service in an intermediary layer, then utilization shortfalls or pricing pressure can cascade quickly. What looks like a stable stream of contracted revenue can become fragile if it depends on continued refinancing, continued access to cheap capital, or continued enthusiasm for long-horizon growth assumptions.

A fourth reason is narrative memory. People remember that previous bubbles often involved financial engineering that was not inherently fraudulent but was brittle: vendor financing, aggressive revenue recognition, off-balance-sheet structures, or securitization chains that behaved well until sentiment shifted. The fear is less “someone is lying” and more “everyone is rationally optimistic at the same time, using structures that amplify downside when optimism breaks.”

Importantly, “circular” does not automatically mean “bad.” In capital-intensive platform markets, strategic financing can be a normal way to accelerate ecosystem growth. The mainstream concern arises when circularity serves as a substitute for organic demand, or when it hides how dependent the system is on continued capital availability and continued high utilization.

1.3 What this paper does and does not claim

This paper is written to be useful to readers who are tired of two unhelpful extremes: breathless inevitability and cynical dismissal.

What this paper does.

It provides a clear definition of circularity and a taxonomy of deal patterns, with an emphasis on cash-flow direction, counterparty roles, and contract structure. It explains why rational firms enter these arrangements, and why scarcity and speed make the loop an efficient solution even without any intention to deceive. It then maps plausible fragility points: utilization gaps, pricing compression, refinancing risk, collateral repricing, and the knock-on effects of a slowdown in capex. Finally, it frames consolidation as a distinct outcome from “crash”: the sector can keep growing while still squeezing out weaker players, and the mechanisms for that squeeze may be distribution and bundling rather than model quality alone.

It also proposes practical stress tests and observable signals that help distinguish three scenarios: stable industrialization, slow deflation, and fast credit unwind. The aim is to replace vibes with simple questions: What utilization is required for the financing stack to hold? How

sensitive are margins to inference pricing? Where are the refinancing walls? How concentrated are counterparties? What happens if compute becomes cheaper faster than expected?

What this paper does not do.

It does not claim that circular deals are fraud, nor does it assert wrongdoing by any specific firm. The point is structural, not accusatory. It does not predict a crash date, because timing is the least reliable part of bubble discourse. It does not assume that model capability progress stops; even fast technical progress can coexist with poor returns and financial fragility. It does not argue that one company "must" win AI across all domains. The paper treats "AI" as multiple markets with different moats: consumer distribution, enterprise procurement, developer platforms, and specialized vertical applications.

It also does not argue that the correct response is to freeze AI development. The question is not whether to build. The question is how to build without confusing financed demand with durable demand, and how to prevent a credit-style unwind from becoming an economy-wide event.

The paper's core claim is narrower: when AI becomes an infrastructure business, the relevant risks and competitive outcomes look like infrastructure businesses. Circular financing and platform interdependence can be incentive-consistent and still be brittle. The task is to identify where brittleness accumulates, what would falsify the crash narrative, and what would confirm it early enough for decision-makers to adapt.

2. The AI Capital Stack in 2024–2026

If the first era of consumer internet felt "weightless," the current AI era is heavy. Its limiting factors are not only models and talent, but also physical throughput: accelerators, networking gear, data-center shells, power delivery, and cooling. That reality produces a recognizable capital stack with multiple layers of risk transfer. The crucial point is that "AI spending" is not one thing. It is a chain of purchases, leases, reservations, and buildouts that convert uncertain future demand into present-day contractual obligations. Those obligations can be distributed across hyperscalers, chip suppliers, specialist infrastructure firms, landlords, utilities, and financiers. Understanding who holds which obligations is the difference between a normal capex cycle and a fragile one.

2.1 Who pays for what: chips, cloud, data centers, power, networks

At a high level, the AI capital stack routes money into five buckets: compute hardware, cloud services, physical data centers, power, and networks. Each bucket has its own buyers, financing tools, and failure modes.

Compute hardware starts with accelerators and the systems that host them. The immediate payers vary:

- Hyperscalers buy large volumes of accelerators and deploy them into their own fleets. They absorb much of the capex and seek to monetize it through cloud services, platform fees, and enterprise contracts.
- Specialist infrastructure providers buy or lease accelerators to resell compute to AI developers and enterprises. This group often finances equipment more aggressively, because its business model is explicitly "turn capex into contracted cash flow."
- Some frontier AI developers buy hardware directly, but many prefer not to. Hardware ownership can be strategically attractive, yet it also concentrates depreciation risk and operational burden.

Cloud spend is the second bucket: compute as a service, storage, managed networking, orchestration, and enterprise security. Here the payer is often the AI developer or enterprise user, but the true economic payer can be an investor or strategic partner who funds the developer, after which the developer spends heavily on cloud. That is one of the simplest pathways by which cash begins to circulate in loops.

Data centers are the third bucket: the shell, racks, power distribution, cooling systems, fire suppression, and the specialized design needed for dense AI loads. The payers here are frequently not the AI developers themselves:

- Colocation providers and data-center landlords finance and build facilities, then lease space and power capacity to hyperscalers, neoclouds, and large enterprises.
- Hyperscalers also build their own facilities, particularly when they want control over cost, integration, and strategic capacity.
- Construction and equipment vendors are paid upfront, while the ultimate economics depend on the long-term lease or service contract that fills the building.

Power is the fourth bucket, and it has become the quiet governor of AI expansion. Dense compute turns electricity into a primary input, not a background utility. Power spending shows up as:

- Long-term electricity supply contracts (or equivalents) that stabilize price and secure availability.
- Grid interconnection fees and the timelines required for new capacity to connect.

- On-site investment in transformers, substations, backup generation, and increasingly sophisticated load management.

Networks are the fifth bucket: high-speed interconnect inside clusters, data-center switching, and wide-area connectivity. Training and serving at scale requires low-latency fabrics and high throughput. Networking capex is a first-order line item, not an accessory. This also means that "AI spend" is not merely GPU spend; it is GPU plus the infrastructure that prevents GPUs from waiting on data.

A useful simplification is to separate capex holders from revenue holders. In an ideal world, the entity that pays for infrastructure also has diversified revenue streams to support it through demand volatility. In a fragile world, the capex is held by entities whose revenue is concentrated, cyclical, and dependent on a small set of counterparties. Much of the current concern about AI finance is simply concern about where the capex has landed.

2.2 The role of long-dated commitments and capacity reservations

In a scarce, time-sensitive market, short-term spot buying is a losing strategy. Many actors therefore reach for long-dated commitments. These commitments are not exotic; they are the standard tools used to make capital-intensive buildouts financeable. What is new is their scale and how central they have become to the AI race.

Long-dated commitments appear in several forms:

- Capacity reservations for cloud compute, including minimum spend arrangements and multi-year purchasing commitments.
- Long-term leases for data-center space and power capacity.
- Equipment leases and financing agreements for accelerators and networking gear.
- "Take-or-pay" style terms in which the buyer pays even if usage falls short of expectations, sometimes with limited flexibility to shift workloads.

These structures solve real problems.

They solve the supplier's problem by making demand legible. A supplier can justify manufacturing allocations, data-center construction, and power interconnections when a customer signs multi-year commitments. These contracts convert uncertain growth narratives into bankable cash-flow expectations.

They solve the buyer's problem by securing scarce inputs and reducing strategic uncertainty. In an arms-race environment, the fear is not overpaying this quarter; it is being capacity-constrained next year while a competitor scales.

They also solve the financier's problem. Debt and structured credit are far easier to issue when there is contracted revenue. A portfolio of long-dated commitments can be presented as stable cash flow, even when the underlying demand is still maturing.

But the same structures create the classic failure mode of infrastructure cycles: utilization mismatch. The commitments are fixed; the demand is not. If revenue growth lags, the buyer can face a squeeze where it must service commitments from a thinner operating margin. If compute prices compress faster than expected, the buyer may be locked into above-market rates or may be forced into renegotiations that damage the supplier's projected cash flows. In extreme cases, commitments that were used to justify new buildouts can turn into stranded capacity.

Long-dated commitments also concentrate counterparty risk. If a large share of a provider's contracted revenue is tied to a handful of fast-growing but still-loss-making AI firms, the provider's "stability" is only as stable as those firms' fundraising environment and customer traction. Conversely, if an AI developer's commitments are mostly to one provider, switching costs rise and bargaining power shifts.

A key diagnostic question for this era is therefore simple: which commitments are genuine hedges against scarcity, and which are substitutes for organic demand? Both can be rational. Only one is robust under disappointment.

2.3 The emergence of the neocloud layer

Between hyperscalers and AI developers, a distinct intermediary class has expanded: specialist GPU cloud providers and infrastructure aggregators, sometimes called the neocloud layer. They exist because the market has two simultaneous truths:

- Hyperscalers cannot instantly meet every surge in demand, and they prioritize their own platform strategy.
- Frontier AI developers and fast-moving enterprises often want large amounts of compute quickly, with terms and operational patterns that hyperscalers may not optimize for by default.

Neocloud providers attempt to fill this gap by doing three things:

First, they aggregate hardware. They acquire accelerators and high-speed networking in volume, often through supply relationships that smaller buyers cannot access.

Second, they aggregate facilities and power. They contract with data-center landlords, secure power capacity, and stand up deployments with a speed that can be competitive in a permit- and grid-constrained world.

Third, they aggregate financing. They turn “we will need GPUs” into “we have GPUs,” using debt, leases, and other credit structures that are justified by contracted customer commitments.

This layer can be economically productive. It can increase overall capacity, improve utilization by pooling customers with different demand profiles, and reduce time-to-compute for developers who would otherwise be stuck in procurement queues.

It also introduces new fragilities:

- Leverage sensitivity. Because neocloud economics often rely on financing hardware and facilities, a tightening of credit conditions can quickly change the viability of expansion plans.
- Collateral repricing. The value of accelerators is not purely intrinsic; it depends on demand and on the pace at which new generations make older hardware less desirable. If the resale market weakens, lenders may tighten terms or require more equity.
- Customer concentration. A neocloud provider may rely on a small number of large AI customers. If one customer renegotiates, delays ramp, or loses funding, the provider’s cash flow can be disrupted abruptly.
- Pass-through risk. If the neocloud’s cost of capital rises but its customers demand lower prices as compute commoditizes, margins can compress from both ends.

The neocloud layer is also where “circularity” becomes most visible. A chip supplier may support the growth of infrastructure intermediaries, who in turn buy more hardware and sell compute to AI developers, who may themselves be funded by strategic investors tied to the same ecosystem. None of that requires deception. It is simply what happens when scarce inputs, platform strategy, and financing machinery converge.

The important analytical move is to stop treating “AI investment” as a single story and to start treating it as a stack. In that stack, different actors carry different kinds of risk: technology risk, utilization risk, credit risk, and market-structure risk. The next sections build on this map to define circularity precisely and to show where brittleness accumulates when growth assumptions meet contractual reality.

3. Defining Circularity

“Circular deals” is a phrase that can mean anything from ordinary platform strategy to genuinely fragile financial engineering. To be analytically useful, it has to be defined in a way that can be recognized in the wild: a set of roles, a direction of cash flows, and a reason the loop matters.

This section therefore treats circularity as a structural pattern, not as an accusation. The goal is to distinguish arrangements that are simply efficient ecosystem-building from arrangements that may amplify fragility by making demand look broader, safer, or more independent than it really is.

A clean working definition is:

Circularity exists when the same economic actors appear in multiple roles across a deal chain such that capital injected on one side returns, directly or indirectly, as revenue on the other side, and the viability of the chain depends materially on that return flow continuing.

That definition does not imply illegality. It implies dependence.

3.1 Circular financing vs normal strategic partnerships

Strategic partnerships are normal in technology markets. Large platforms invest in suppliers, suppliers invest in customers, and everyone signs multi-year agreements to reduce uncertainty. Many of these arrangements involve mutual benefit and perfectly healthy incentives.

Circular financing is narrower and more concerning, but only under specific conditions.

A normal strategic partnership typically has these features:

- The investment has an independent rationale that would survive even if the partner purchased less in the short term.
- The commercial relationship has market-like terms that could plausibly be replaced by third parties.
- The investor is not relying on the partner's purchases to justify the investment's near-term economics.
- The partnership improves product integration, distribution, or capability in a way that plausibly creates real end-user value.

Circular financing, by contrast, tends to have one or more of these features:

- The investment and the procurement are coupled in practice, even if not explicitly tied: the capital enables the partner to purchase the investor's goods or services at scale.
- The investment serves as a bridge to meet minimum commitments: it prevents a customer from failing to pay for reserved capacity or leased infrastructure.
- The return flow of spending is large relative to the partner's organic revenue, making demand dependent on continued financing.

- The arrangement increases leverage or increases reliance on refinancing by anchoring debt issuance to contracted payments from a small set of related counterparties.

An important distinction: circularity is not about reciprocity ("we invest and you buy from us"). That is common. Circularity is about *material reliance*: the system's apparent demand or stability depends on money that effectively originates from within the same small network of players.

This is why the debate often sounds moralistic when it should be mechanical. The right question is not "Is this shady?" The right question is "If growth slows, does this loop still hold without self-funding?"

3.2 Round-tripping patterns: investor-supplier-customer loops

Round-tripping is the simplest circular pattern: money leaves an entity and then comes back as revenue, sometimes after passing through one or more intermediaries. In AI infrastructure, the loop commonly has three roles:

- Investor or capital provider: provides equity, credit support, prepayments, or other funding.
- Supplier: sells compute, cloud services, chips, networking, or facility capacity.
- Customer: buys those products, often using the funds raised.

In many AI cases, a single firm can play more than one role. A hyperscaler can be both investor and supplier. A chipmaker can be both supplier and financier (directly or indirectly) through ecosystem support. An AI lab can be both customer (of compute) and supplier (of API services) while being financed by platform partners who also buy its services.

Common loop shapes include:

- Platform-to-lab-to-platform: a platform funds an AI developer; the developer uses the money to buy cloud compute and services from the platform; the platform can then cite increased cloud demand as evidence of momentum.
- Supplier-to-infrastructure-to-lab-to-supplier: a hardware supplier supports an infrastructure aggregator; the aggregator buys more hardware and sells compute to labs; the labs' compute spend helps validate more expansion and financing.
- Multi-party echo loop: several incumbents invest in a frontier lab; the lab signs major capacity commitments; intermediaries finance infrastructure buildout based on those commitments; the same incumbents supply critical inputs and receive much of the downstream spend.

In all cases, the loop can be economically real. GPUs are delivered, power is consumed, and services are provided. The question is whether the "demand signal" is independent enough to predict durable cash flow once the funding environment tightens.

Round-tripping becomes materially risky when:

- The loop is used to maintain growth optics (bookings, commitments, contracted revenue) rather than to support profitable end-user adoption.
- The loop props up an intermediary's debt service, so that one weak link can cause a cascade.
- The loop is reinforced by concentration, meaning there are few alternate counterparties to absorb the shock.

3.3 Where circularity shows up in contracts and cash flows

Circularity often hides in plain sight because it is not a single contract. It is a chain of contracts whose net effect is circular.

Where to look:

- Equity + commercial commits: an investment round is followed by, or coincides with, a large multi-year purchasing commitment to the investor's services or hardware ecosystem.
- Prepayment structures: a customer prepays for compute or capacity, giving the supplier upfront cash that finances expansion, while the customer relies on future usage to justify the prepayment.
- Take-or-pay obligations: the buyer must pay regardless of usage, creating pressure to raise capital or refinance to meet the obligation if utilization lags.
- Most-favored and bundling terms: pricing and integration agreements that make it difficult to move spend elsewhere, effectively locking the loop in place.
- Collateralized equipment finance: accelerators are purchased via leases or loans; lenders underwrite based on contracted customer payments; the customer's ability to pay is tied to continued access to funding or continued high-margin demand.
- Related-party adjacency: not necessarily formal related-party transactions, but networks of investors and partners that function as an ecosystem, with repeated counterparties and synchronized incentives.

From a cash-flow perspective, circularity reveals itself in two ratios:

- Internalized demand ratio: how much of a firm's "demand" is ultimately financed by capital originating from its own platform ecosystem rather than from independent end customers.
- Commitment coverage ratio: how much of a buyer's long-dated commitments are covered by stable, diversified operating cash flow versus by ongoing fundraising.

The contracts can be perfectly legitimate and still produce a fragile system if those ratios become extreme.

3.4 A taxonomy of deal types and their risk signatures

To keep the conversation concrete, it helps to classify deal patterns by what they optimize for and what they break under stress. Below is a practical taxonomy.

Type A: Integration partnerships
Mechanism: investment plus product integration, co-selling, shared roadmaps.
Healthy signature: value creation is visible in end-user adoption; switching remains feasible.
Risk signature: low. Stress impact is mostly competitive, not systemic.

Type B: Capacity assurance deals
Mechanism: multi-year commitments to secure scarce compute, often with minimum spend.
Healthy signature: buyer has diversified revenue; utilization is likely; contract supports genuine scaling.
Risk signature: moderate. Breaks via utilization shortfall or pricing compression.

Type C: Vendor financing analogs
Mechanism: supplier-adjacent funding enables customer purchases (explicitly or effectively).
Healthy signature: funding is small relative to customer's independent revenue; purchases reflect real demand.
Risk signature: elevated. Breaks when the customer cannot fund ongoing purchases without continued support.

Type D: Balance-sheet optics loops
Mechanism: large commitments or prepayments that make demand look stronger, enabling further fundraising or favorable narratives.
Healthy signature: rare. Often hard to justify without significant independent demand.
Risk signature: high. Breaks when the market reprices growth expectations and scrutinizes quality of revenue.

Type E: Intermediary leverage stacks
Mechanism: infrastructure aggregators finance GPUs and facilities with debt justified by contracted payments from a small set of AI customers.
Healthy signature: customer base diversified; refinancing needs modest; collateral resilient.
Risk signature: high. Breaks through credit tightening, customer concentration, collateral repricing, and refinancing walls.

Type F: Ecosystem circularity at scale
Mechanism: multiple incumbents funding, supplying, and distributing across the same set of frontier players; spend circulates across the same nodes.
Healthy signature: ecosystem produces widely adopted products that generate new, independent revenue streams.
Risk signature: ambiguous but potentially systemic. Breaks through synchronized deleveraging or a shock that hits multiple nodes simultaneously.

This taxonomy supports two disciplined conclusions.

First, circularity is not a binary. Some loops are simply how platforms build ecosystems under scarcity. Others are fragile because they substitute financing for demand or because they create leverage-dependent cash-flow assumptions.

Second, the "AI crash" narrative becomes plausible only when multiple high-risk types cluster: vendor-financing analogs plus leveraged intermediaries plus long-dated commitments whose coverage depends on continued fundraising. The next sections will use this definition and taxonomy to explain why rational actors choose the loop and where fragility hides when the loop becomes the market's default way to scale.

4. Why Rational Actors Choose the Loop

It is tempting to treat circular financing and ecosystem loops as inherently suspect. That temptation usually comes from importing a retail-investor moral story into what is, in practice, a coordination problem under scarcity. Many of these arrangements exist because the alternative is worse: missing compute windows, losing strategic position, or failing to translate model progress into distribution before the market consolidates. In this sense, the loop is not a trick. It is a mechanism for converting uncertainty into capacity quickly. The question is not why anyone would do it. The question is what assumptions must remain true for it to stay safe.

Four forces make the loop rational: scarcity, urgency, distribution gravity, and narrative incentives.

4.1 Scarcity economics: GPUs, power, permitting, and time-to-ship

The fastest way to understand why loops arise is to notice what is scarce. The bottlenecks of 2024–2026 are not abstract. They are tangible and slow-moving.

GPUs and high-end accelerators are scarce because advanced manufacturing is capacity-constrained, the hardware is complex, and supply is concentrated. Even when production expands, allocation tends to favor the largest buyers with the most predictable order volumes and the deepest relationships. That reality pushes smaller or newer buyers toward intermediaries, and it pushes everyone toward commitments that justify priority access.

Power is scarce in the only sense that matters: reliable, deliverable megawatts at the right location and timeline are hard to secure. Many regions face multi-year queues for grid interconnection upgrades. AI loads are dense and continuous. Utilities must plan, approve, and build. Substations, transformers, and transmission lines are not instantaneous. If compute is the brain of AI, power is its metabolism, and metabolism constraints are unforgiving.

Permitting is scarce because land use, environmental review, zoning, and public opposition impose delays. Data centers are not invisible. They involve construction traffic, water use, noise, and the perception of "industrialization" in communities that may not want it. Even when approvals are likely, they take time. Time-to-permit is a competitive variable.

Time-to-ship is scarce because the whole supply chain has synchronized lead times: chips, servers, racks, cooling equipment, switchgear, networking, and construction scheduling. Any delay in one component can stall the entire deployment. In that environment, waiting for purely organic demand pull is strategically risky, because by the time demand is obvious, capacity may be unavailable.

Scarcity creates loops because loops are coordination solutions. A party with capital and access to scarce inputs can fund a party with demand and technical capability, and the funded party can direct spend into the scarce input channels. This turns "potential demand" into "credible demand," which triggers suppliers and builders to commit resources. The loop can therefore be understood as a scarcity-to-capacity conversion process.

The fragility comes from mistaking scarcity conversion for market validation. Scarcity can force demand into certain channels even if end-user willingness to pay is not yet proven at scale.

4.2 Competitive urgency: spending as a defensive moat

The second force is urgency. In a fast-moving technological race, there is a window in which compute becomes a strategic weapon rather than a commodity. During that window, spending can function as a defensive moat in several ways.

First, scale can accelerate learning. More compute enables more experiments, more training runs, more data synthesis, more fine-tuning, more evaluation, and faster iteration. This can create real capability advantages, especially when product cycles are measured in weeks.

Second, spend secures optionality. A firm that reserves capacity and builds infrastructure can pivot among workloads: training, inference, fine-tuning, and internal tools. A firm that does not reserve capacity becomes a price-taker and a queue-taker.

Third, spend can deny rivals. In constrained markets, large commitments do not only secure supply; they reduce what is available to competitors. Even if that is not the explicit intent, the effect can be similar.

Fourth, spend can stabilize partnerships. Long commitments can lock in co-development, preferential access, and integration advantages. This matters when the ecosystem is still selecting its winners.

In an arms race, what looks like overbuilding from a static viewpoint can be rational from a dynamic one. The expected loss from being late is higher than the expected loss from overpaying. This is why rational actors will sign large commitments before returns are clear: the dominant fear is strategic displacement.

The crash risk arises when "defensive moat spending" persists after the moat has stopped providing marginal advantage. Once models and inference become more commoditized, excessive fixed commitments turn from moat to millstone.

4.3 Distribution advantages: default surfaces and bundled adoption

The third force is distribution, and it is perhaps the most underestimated. Many readers implicitly assume that the "best model" wins. In practice, "best model" is only one input into adoption. What often matters more is who controls default surfaces: the places where users already are.

Default surfaces include search, browsers, mobile operating systems, app stores, productivity suites, enterprise identity systems, and cloud marketplaces. If a firm can place AI features into these surfaces, adoption can be bundled, frictionless, and hard to dislodge. This creates a platform advantage that can outlast short-term model quality differences.

This distribution gravity encourages loops because incumbents can finance ecosystem growth in ways that increase their own distribution capture.

A platform may invest in an AI developer, not only to benefit from financial upside, but to ensure that the developer's product fits tightly into the platform's distribution surfaces. The developer then spends on the platform's cloud, strengthening the platform's infrastructure

economics. The platform then ships AI features into its default products, monetizing through existing channels. The loop is coherent: capital, compute, and distribution reinforce each other.

This dynamic also supports the “squeeze” thesis. Even if many competitors build excellent models, default-surface incumbents may still capture the majority of user attention and monetization. That does not require eliminating competitors. It requires making competitors marginal, niche, or dependent.

The structural risk is that the loop can make the ecosystem look more pluralistic than it is. Many logos can exist at the application layer while a small number of platforms quietly control the bottlenecks: distribution, identity, billing, and the compute substrate.

4.4 Narrative arbitrage: valuation support and “proof of demand” optics

The fourth force is narrative, and it is the most uncomfortable to discuss because it sits at the border of psychology and finance. Yet it is central to why loops can grow large.

Narrative arbitrage happens when a deal structure produces a story the market rewards, even if the underlying economics are still uncertain. In capital markets, stories matter because they influence the cost of capital, and the cost of capital determines who can build.

The current AI environment rewards several narratives:

- “Demand is exploding.” This is supported by large commitments, booked revenue, and announcements of capacity expansion.
- “We have secured supply.” This is supported by multi-year reservations, hardware allocations, and data-center buildouts.
- “We are the platform.” This is supported by ecosystem partnerships, integrations, and the appearance of indispensability.

Circular structures can strengthen these narratives by making demand more legible and more headline-friendly. A major purchasing commitment looks like validation. A multi-billion-dollar data-center expansion looks like inevitability. An investment round with famous strategic backers looks like endorsement.

None of this requires deception. But it can produce a feedback loop: capital funds commitments, commitments justify expansion, expansion sustains narrative, narrative lowers cost of capital, lower cost of capital funds more commitments.

This is the key mechanism by which “proof of demand” optics can become self-referential. In a normal market, demand proof is end-user willingness to pay. In a scarcity-driven platform

market, demand proof can be replaced by contractual commitments financed by the same network that benefits from the commitments.

The risk signature of narrative arbitrage is not that the assets are fake. It is that the valuation assumes a smooth translation from commitments to profitable utilization, and from utilization to durable pricing power. If either assumption fails, the same narrative machinery can reverse, increasing the cost of capital precisely when refinancing is needed.

The disciplined takeaway is that rational actors choose the loop because it solves immediate coordination problems under scarcity, urgency, and distribution pressure, while also aligning with the narratives that lower capital costs. The next section identifies where fragility hides inside these rational choices, and why the system can be both sensible in the short term and brittle in the long term.

5. Where Fragility Hides

The danger in the current AI boom is not that it is “unreal.” The assets are physical. The chips exist, the power is consumed, the data centers are built, and contracts are signed. Fragility hides somewhere else: in the coupling between fixed obligations and uncertain cash flows, in the way credit markets translate optimism into leverage, and in the way reporting conventions can turn a slowdown into a sudden repricing event. When the system is healthy, these structures look like prudent scaling. When conditions change, the same structures can amplify shocks.

This section identifies four fragility channels that can coexist with perfectly rational behavior.

5.1 Leverage and maturity mismatch in the financing chain

Infrastructure booms rarely fail because someone “ran out of ideas.” They fail because someone ran out of refinancing options. That is the core leverage risk: borrowing short to fund long-lived assets whose cash flows are still maturing.

In the AI capital stack, leverage appears at multiple points:

- Hardware financing: accelerators and networking gear are acquired via leases or secured loans that assume steady utilization and stable secondary-market value.
- Data-center financing: landlords and developers finance construction with debt that is repaid by long-term lease payments and power-fee streams.
- Intermediary financing: infrastructure aggregators often rely on debt backed by contracted customer payments, with covenants tied to utilization, concentration, or liquidity.

Leverage by itself is not pathological. It is what makes large-scale buildouts possible. The fragility emerges when maturities do not match.

Maturity mismatch looks like this:

- Assets have long economic lives, but the financing has shorter rollover dates.
- Customer contracts may be long-dated on paper, but they include renegotiation triggers, termination options, or usage-dependent pricing that weakens cash-flow certainty.
- The borrower's business model assumes that it can refinance at similar or better terms, which is only true when credit is abundant and confidence is high.

When conditions tighten, maturity mismatch becomes a cliff. Even if the underlying business is viable in the long run, a near-term refinancing wall can force distressed behavior: asset sales, dilutive fundraising, emergency renegotiations, or, in the worst case, default.

This is where circularity matters: if multiple layers depend on the same small set of counterparties continuing to spend, and those counterparties depend on continued fundraising, then a credit tightening can propagate quickly. The system can become synchronized: many firms attempt to reduce commitments and conserve cash at the same time, precisely when lenders become less willing to roll debt.

A practical red flag is leverage stacked on leverage: the end demand is still uncertain, but the chain from chips to compute to contracts to credit has multiple points where refinancing is assumed rather than guaranteed.

5.2 Utilization risk: take-or-pay, minimum commits, and ramp schedules

Most "AI crash" stories become concrete when you talk about utilization. Utilization is the bridge between capex and revenue. High utilization turns fixed infrastructure into efficient economics. Low utilization turns fixed infrastructure into dead weight.

The main utilization traps in 2024–2026 are contractual, not technical.

Take-or-pay terms require payment regardless of usage. These terms appear in different disguises: minimum spend, reserved capacity that expires, fixed lease payments, power-capacity charges, and penalties for underuse. They exist because suppliers need bankable demand to justify investment. From the buyer's perspective, they are the price of securing scarce resources.

Minimum commitments become risky when revenue is back-end loaded. Many AI businesses have a ramp schedule: they expect adoption to scale over time, so commitments are signed

today based on anticipated demand tomorrow. If the ramp slips, the buyer can face a squeeze: paying for capacity while still searching for profitable customers.

Ramp schedules are fragile because they rely on multiple dependencies:

- Product adoption must grow fast enough.
- Pricing must remain strong enough.
- Costs per unit of inference must fall at the right rate.
- Competition must not force a rapid price war.
- Financing must remain available if early cash flow is insufficient.

A small miss in each dependency can multiply into a large miss overall. That is why utilization risk can feel sudden. A business can be "on plan" narratively while drifting off plan financially.

Another subtle utilization issue is workload mismatch. Training demand is spiky; inference demand can be steadier but price-sensitive. A fleet optimized for training can be underutilized if training cycles slow, and a fleet optimized for inference can suffer if marginal revenue per token collapses. The ability to repurpose capacity across workloads matters, but it is not unlimited.

Utilization fragility becomes systemic when many actors rely on the same story: "utilization will rise to cover commitments." If the sector's growth trajectory is revised downward, underutilization can appear across many nodes at once, stressing lenders, landlords, and suppliers simultaneously.

5.3 Collateral dynamics: what a GPU is worth in a downturn

In equipment-financed booms, collateral is the silent stabilizer. It reassures lenders that, if a borrower fails, the asset can be sold. That reassurance is only as good as the secondary market.

A GPU's value in a downturn is not like a truck's value in a downturn. It is more volatile, for three reasons.

First, technological obsolescence is faster. New generations can shift performance-per-watt and performance-per-dollar quickly, reducing demand for older units. Even if older GPUs remain useful, the premium buyers pay can compress sharply when better options are available.

Second, the buyer pool is concentrated. The secondary market is not a broad, liquid market with millions of small buyers. It is a market dominated by a few large players and a limited set of specialist operators. In a downturn, those buyers may step back at the same time, reducing liquidity.

Third, collateral value depends on complementary constraints. A GPU is not valuable in isolation. It needs power, cooling, networking, and a place to run. If data-center capacity and power interconnections are also constrained, that can support GPU value. But if the downturn includes a pause in new buildouts or a glut in certain facilities, the effective value of hardware can become location- and integration-dependent.

In a credit-driven expansion, lenders may assume relatively stable collateral value. If that assumption breaks, lending terms tighten, margin requirements rise, and refinancing becomes harder. This can cause a feedback loop: falling collateral values lead to tighter credit, which forces sales, which further depresses prices.

A practical way to express this is: collateral is pro-cyclical. It looks strong when everyone wants to expand. It looks weak when everyone wants to shrink. The more the capital stack relies on equipment finance, the more sensitive it becomes to that pro-cyclicality.

5.4 Accounting and reporting pressure points that can amplify shocks

Even in perfectly legitimate markets, reporting conventions can amplify shocks by shaping how investors perceive the quality of growth. The problem is not that reporting is "fake." The problem is that certain metrics can look strong even when the underlying economics are weakening.

Pressure points include:

- Commitments versus realized usage. Large contracted commitments can be interpreted as durable demand, but they may not translate into profitable utilization if usage is below contracted levels or if pricing is pressured.
- Revenue recognition timing. Different contract structures can shift when revenue appears on financial statements relative to when cash is received and relative to when costs are incurred.
- Gross versus net economics. A company can show headline revenue growth while still subsidizing usage through credits, discounts, or bundling, leaving unit economics unclear.
- Capitalization choices. The line between operating expense and capital investment can influence reported profitability and can mask the ongoing cost of maintaining and upgrading infrastructure.
- Related-ecosystem dependence. When counterparties are entangled—investors, suppliers, and customers in overlapping networks—headline growth can look diversified while actually being concentrated.

The amplification mechanism is straightforward:

1. In the boom phase, markets reward growth signals such as commitments, capacity expansion, and contract announcements.
2. As skepticism rises, markets demand proof of cash flow, utilization, and margin durability.
3. If the proof is weaker than the commitments implied, repricing can be abrupt.
4. Repricing raises the cost of capital, tightening refinancing options exactly when leverage is highest.

This is how "sentiment" becomes structural. A shift in what investors consider credible can change financing conditions, which changes the real economy of the AI stack.

The central message of this section is not pessimism. It is specificity. Fragility is not a vibe; it is a set of mechanisms:

- leverage that assumes refinancing,
- commitments that assume utilization ramps,
- collateral that assumes liquid resale value,
- reporting that can overemphasize commitments over cash flow.

If those mechanisms remain aligned—utilization rises, pricing holds, refinancing stays available—then the system can industrialize successfully. If they become misaligned, the same rational structures that enabled rapid scaling can become the channels through which the boom unwinds.

6. Crash Pathways

A useful way to talk about an "AI crash" is to stop treating it as a single event and instead treat it as a set of pathways that share one common feature: fixed obligations meet weaker-than-expected cash flow. The pathways differ mainly by speed, trigger, and where the first visible break occurs. Some scenarios look like a gentle deflation of exuberance. Others look like a credit event that forces rapid repricing and liquidation. In practice, these pathways can overlap: a slow ROI disappointment can set the stage for a fast crash once credit conditions tighten.

6.1 The slow crash: ROI disappointment and capex deceleration

The slow crash is the most common “bubble” outcome in infrastructure cycles: the story does not end with a dramatic collapse, but with a long period of underperformance, consolidation, and write-downs. The mechanism is simple. AI adoption continues, but the returns do not justify the pace of buildout.

The sequence typically looks like this:

First, enterprises and consumers adopt AI unevenly. Many deployments prove useful, but not transformative enough to support premium pricing at the scale implied by peak investment narratives. Internal productivity gains appear, yet they are hard to monetize directly. External products grow, yet competition compresses price. The result is that revenue grows slower than the most aggressive infrastructure assumptions.

Second, unit economics drift. Even if usage rises, inference pricing can fall faster than costs decline, especially when competitors treat AI as a feature rather than a standalone profit center. If AI is bundled into suites, the marginal revenue per unit of compute can become opaque, and compute-heavy features may be justified for retention rather than for incremental margin.

Third, capital discipline returns. Boards and investors begin asking narrower questions: Which workloads are profitable now? Which customers are paying real cash rather than credits? Which commitments are truly needed versus defensive hoarding? This is where capex deceleration begins. Expansion plans are delayed, data-center pipeline is re-scoped, and the growth of reserved capacity commitments slows.

Fourth, the ecosystem reprices growth expectations. Equity valuations compress, but operating activity does not stop. The crash, in this pathway, is largely a valuation and investment-rate crash rather than a collapse of usage.

In this scenario, the primary casualties are not the largest platforms. They are the marginal players whose business model assumes high growth and abundant capital: highly levered infrastructure intermediaries, single-product AI firms with large fixed commitments, and overbuilt data-center projects in less advantaged locations.

Early indicators for the slow crash include:

- rising skepticism around ROI claims and longer enterprise sales cycles
- a shift from “capacity announcements” to “efficiency and utilization” messaging
- cancellations, deferrals, or quiet renegotiations of expansion plans
- consolidation framed as “synergies” rather than “growth”

6.2 The fast crash: credit tightening and forced deleveraging

The fast crash is a financing event. It happens when the cost of capital rises or liquidity dries up while the sector still depends on refinancing to sustain expansion or even to maintain existing obligations.

The trigger can come from many places: a macro rate shock, a credit event in a neighboring sector, a sudden risk-off move in private markets, a regulatory shock affecting disclosure or underwriting, or a single high-profile default that causes lenders to reassess assumptions across the entire AI infrastructure complex.

The sequence tends to be sharper:

First, lenders tighten. Terms worsen: higher spreads, stricter covenants, more collateral requirements, less willingness to roll debt, and more scrutiny of customer concentration and contract quality.

Second, refinancing walls appear. Firms that assumed they could roll short- or medium-term financing face a gap between what they need and what the market will offer. If they cannot raise equity at acceptable terms, they are pushed toward emergency measures.

Third, forced deleveraging begins. Assets are sold, contracts are renegotiated, capex is halted, and expansion projects are paused midstream. Importantly, this can happen even if demand is still growing, because the constraint is financing, not usage.

Fourth, counterparty fear spreads. When one node becomes distressed, other nodes question whether their counterparties will meet obligations. This can lead to cascading conservatism: everyone tries to reduce risk exposure at the same time.

The fast crash is where “circularity” becomes dangerous. If a material fraction of demand is financed by the same network of actors, then a tightening in one part of the network reduces spend elsewhere, which tightens conditions again. The loop that was stabilizing in the boom phase becomes destabilizing in the bust phase.

Early indicators for the fast crash include:

- widening credit spreads or tightened lending terms for infrastructure-like borrowers
- increased emphasis on liquidity, runway, and covenant headroom
- sudden freezes in new equipment financing or new data-center project finance
- distressed renegotiations framed as “strategic realignment”

6.3 Secondary-market cascades: price discovery for compute and hardware

The slow crash is about disappointing returns; the fast crash is about credit. The secondary-market cascade is about price discovery. It is what happens when the market is forced to decide, quickly, what compute and hardware are actually worth when everyone is no longer expanding simultaneously.

In the boom, compute is often priced by scarcity and urgency. In the bust, compute is priced by surplus and bargaining power.

The cascade can unfold in layers:

First, compute pricing compresses. Buyers demand better terms, more flexibility, and shorter commitments. Sellers discount to maintain utilization. Price cuts may be framed as "efficiency gains," but the effect is margin compression for intermediaries.

Second, hardware resale markets loosen. If expansion slows, fewer buyers need incremental GPUs immediately. If new generations arrive, older units lose premium status. If multiple sellers hit the market at once, liquidity evaporates and prices gap down.

Third, collateral assumptions reset. Lenders mark collateral down, increase haircuts, and reduce willingness to underwrite new purchases. This reinforces the downturn by restricting the financing channel that enabled rapid growth.

Fourth, a feedback loop emerges: lower compute prices reduce revenue per unit, which pressures utilization economics, which pushes more sellers into the market, which further lowers prices.

This is the pathway where observers start talking about "the real clearing price" of AI. But the key point is that there may not be a single clearing price. Compute is not perfectly fungible. Latency, location, networking, power availability, and software integration all matter. Some capacity remains valuable; other capacity becomes stranded.

Early indicators for secondary-market cascades include:

- aggressive discounting for new capacity commitments
- rising churn or shorter-term contracting preference among buyers
- increased reporting of used hardware availability or "excess capacity"
- lenders tightening collateral terms even without a headline default

6.4 Spillovers: real estate, power markets, and local finance exposure

The most important question is not whether AI players suffer in a downturn. The question is whether the downturn leaks into adjacent sectors. Spillovers depend on how broadly the AI buildout has been embedded into local economies and financing systems.

Real estate spillovers can take several forms:

- data-center land and facility values can reprice if utilization assumptions fall
- construction pipelines can stall, impacting contractors, suppliers, and local employment
- specialized industrial builds can be harder to repurpose than optimistic investors assume

Power market spillovers are subtle but real:

- utilities and grid operators may have planned upgrades based on expected load growth
- long-term power capacity arrangements can become politically sensitive if communities perceive “overbuild for outsiders”
- local ratepayers can become a political focal point if infrastructure costs were socialized while benefits appear privatized

Local finance exposure is the channel that turns a sectoral correction into a regional stress:

- municipalities may have offered incentives or financed infrastructure improvements expecting sustained growth
- local banks and lenders may have indirect exposure through construction loans, real estate loans, or vendor relationships
- concentrated local dependence on a few large projects can create vulnerability if those projects pause

Spillovers do not require an AI apocalypse. They require only a mismatch between investment plans and realized ramp. A region that built for a multi-year surge may face a multi-year digest instead.

The disciplined conclusion of this section is that “AI crash” is not one prediction; it is a menu of mechanisms. The slow crash is a long under-delivery of returns relative to investment. The fast crash is a credit tightening that forces deleveraging. The secondary-market cascade is a rapid repricing of compute and collateral. Spillovers are the transfer of stress into real estate, power systems, and local finance. The next section examines the other major outcome that can occur even without a crash: consolidation, where the system survives but fewer players do.

7. Consolidation Thesis

“Crash” is only one way an AI boom can end. Another way is consolidation: the system keeps growing, but the set of firms that capture most of the value shrinks. In this outcome, there is no single dramatic failure. Instead, the market quietly sorts into a few dominant platforms plus a long tail of dependent suppliers, niche specialists, and feature vendors. Circular financing and long-dated commitments can still matter here, but the central mechanism is not credit. It is control of chokepoints: distribution surfaces, identity and billing rails, procurement pathways, and the compute substrate.

The consolidation thesis matters because it can coexist with both optimism and disappointment. AI capability can advance quickly while returns concentrate in incumbents. The ecosystem can look vibrant at the application layer while power consolidates underneath. If this thesis is right, then many companies will not be “beaten on models.” They will be squeezed by economics and access.

7.1 How incumbents can squeeze the field without “winning the models”

Most people narrate AI competition as a race for the best model. That is a partial truth. In practice, incumbents can squeeze the field through mechanisms that do not require them to permanently lead on raw model quality.

One mechanism is default distribution. The incumbent that controls where users already are can make AI a background feature rather than a separate destination. If AI is integrated into search boxes, browsers, operating systems, email clients, document editors, and enterprise suites, then the marginal value of a standalone AI brand declines. Users may still admire the “best” model, but they will often accept the “good enough” model embedded into their daily workflow.

A second mechanism is cross-subsidy. Incumbents with large profit pools can treat AI as a retention lever, a bundle enhancer, or a defensive cost center. They can price AI below standalone competitors’ cost structure, not because the AI itself is profitable, but because it protects a larger business line. A challenger must earn margin from AI; an incumbent may only need AI to reduce churn or expand engagement elsewhere.

A third mechanism is contracting power. Large platforms can offer integrated enterprise deals that combine cloud credits, security tooling, compliance features, support, and AI usage under one commercial umbrella. This packaging reduces procurement friction and shifts negotiations from “Is this model best?” to “Is this a safe vendor with predictable terms?” Standalone AI companies often struggle not on features, but on procurement and governance overhead.

A fourth mechanism is product surface capture. In many categories, the UI is the moat. If an incumbent controls the surfaces where AI is invoked, it can shape user habits and define what counts as normal. Once habits form, competitor offerings feel like extra steps. This is especially important for AI, where ease of invocation and context integration often matter more than marginal improvements in answer quality.

A fifth mechanism is standard-setting and compatibility gravity. Incumbents can influence toolchains, APIs, evaluation norms, enterprise integration patterns, and deployment architectures. Even if they do not monopolize standards, they can make their stack the default "safe choice." Over time, ecosystems accrete around defaults.

A sixth mechanism is data and feedback loops, but not in the naive sense of "more data always wins." The more relevant form is product telemetry and workflow embedding. Incumbents can see where AI helps and where it fails inside real, high-frequency workflows. They can iterate not only on the model, but on the product system around the model: prompting scaffolds, retrieval, guardrails, latency optimization, and UI design. This can produce compounding advantage even when base models are competitively close.

A seventh mechanism is cost and capacity control. When compute is scarce or expensive, the firm that controls allocation can choose who gets to scale. Even if the incumbent sells compute to competitors, it sets terms and can prioritize its own internal workloads when constrained.

These mechanisms define "squeeze without winning the models." The market can remain competitive in model research while consolidating in distribution, procurement, and unit economics.

7.2 Google's potential choke points: search, ads, Chrome, Android, workspace

The "Google squeeze" argument is a specific instance of the consolidation thesis. It does not require Google to permanently dominate model quality. It requires Google to convert existing control over daily digital entry points into AI defaults, and then to monetize AI through the advertising and productivity ecosystems it already operates.

Search is the most obvious choke point. If the primary interface to information becomes AI-mediated inside the search experience, then AI usage becomes a feature of search rather than a separate product category. That shifts competition from "who has the best chatbot" to "who controls the query funnel." The winner in that funnel can decide what AI results look like, what sources are surfaced, and how commercial intent is routed.

Advertising is the monetization choke point. Google's ad system is not just a revenue line; it is a mature mechanism for pricing attention and commercial intent. If AI becomes a new presentation layer for intent, the key question becomes: who captures the intent and who runs

the auction. A standalone AI assistant might attract users, but if it cannot monetize intent as efficiently, it may remain structurally disadvantaged.

Chrome is a surface choke point. Browser-level integration can make AI assistance available across the web, not only within one app. It can also influence defaults: where users type, what they see first, which services are invoked automatically, and how authentication and permissions are handled.

Android is another surface choke point. Operating systems can integrate AI into notifications, voice, messaging, settings, and system-level tasks. An AI assistant at the OS layer can become the “first resort” for many actions. That reduces the need to open a separate AI app, which reduces the growth headroom for independent AI brands.

Workspace is the enterprise seat choke point. Productivity suites control high-value contexts: email, documents, spreadsheets, meetings, calendars, and collaboration. If AI features are bundled into these workflows, adoption becomes automatic at the seat level. For many enterprises, the question becomes not “Which AI model do we buy?” but “Which features are already available in the tools we pay for?” This shifts AI from a procurement category into a line item inside existing licenses.

Google also has infrastructure choke points: its data centers, networking, and custom accelerator strategy. The consolidation thesis does not require any single hardware advantage to be permanent. It requires sufficient capacity control and cost structure to make aggressive bundling sustainable.

The squeeze scenario becomes more plausible when these chokepoints reinforce each other:

- AI inside search increases engagement and keeps intent inside Google surfaces.
- Ads monetize that intent efficiently, funding further investment.
- Chrome and Android make AI invocation habitual across contexts.
- Workspace normalizes AI inside enterprise workflows and governance.
- Infrastructure scale lowers marginal cost and supports bundled pricing.

In this picture, independent AI companies still exist, but many become either distribution-dependent or niche-specialized. The main risk for the ecosystem is not that innovation stops, but that bargaining power and margins concentrate at the platform layer.

7.3 Cloud platform gravity: switching costs and integration bundling

Even if distribution surfaces were neutral, cloud platforms exert gravity. Once AI workloads are built on a cloud stack, it is expensive and risky to move them. This is not just "vendor lock-in" rhetoric; it is a combination of technical friction, organizational inertia, and economic penalties.

Switching costs arise from several sources:

Data gravity. Training pipelines, feature stores, vector databases, logs, and evaluation datasets accumulate where the compute is. Moving large data sets is slow, expensive, and operationally risky.

Operational entanglement. Identity and access management, network configuration, observability tooling, incident response, and security policies become deeply integrated into one provider's patterns. Rebuilding them elsewhere is not merely migration; it is re-architecture.

Performance coupling. Distributed training and high-throughput inference often depend on the provider's specific networking topology, storage performance, and accelerator integration. Portability exists in theory but can be costly in performance and reliability.

Commercial coupling. Discounts, committed-use arrangements, credits, and enterprise agreements create economic penalties for reducing spend. A firm may be technically able to move but financially constrained by the shape of its contracts.

Integration bundling amplifies these effects. Cloud providers can bundle AI services with managed databases, security, analytics, and deployment platforms. The customer buys "a working system," not a pile of components. This reduces friction and risk for the buyer, but it also increases dependency.

This is where the consolidation thesis intersects with circularity. If capital and compute commitments flow through the same dominant platforms, then startups and intermediaries can become structurally dependent on the very platforms they might otherwise challenge. The platform does not need to forbid competition. It only needs to make competition economically narrow and migration inconvenient.

7.4 Counterforces: multi-cloud strategies, open weights, and specialization

Consolidation is not inevitable. The same conditions that favor incumbents also create incentives for counterstrategies. Several counterforces can prevent a full squeeze, or at least preserve meaningful competitive space.

Multi-cloud strategies are the most direct countermeasure to platform gravity. Enterprises can split workloads across providers, keep core data in portable formats, standardize deployment

tooling, and negotiate better terms by maintaining credible alternatives. Multi-cloud is difficult, but it is increasingly treated as a resilience and bargaining strategy rather than as a pure engineering preference.

Open weights and model portability reduce dependence on any one vendor's APIs. If high-quality models are available for self-hosting or through multiple providers, then base-model capability can commoditize faster. This shifts competition toward integration quality, tooling, and operational excellence, and it reduces the risk that one firm controls the only viable model supply.

Specialization creates defensible niches. Many real-world use cases require domain adaptation, data governance, latency constraints, or regulatory compliance that a general-purpose platform may not optimize for. Vertical AI companies can win by owning domain workflows, trusted data pipelines, and specialized evaluation standards. They can also win by being closer to the customer's operational loop, where value is measured in outcomes rather than in benchmarks.

On-prem and edge inference are another counterforce. If inference becomes efficient enough to run on enterprise hardware or local devices for many tasks, then the power of centralized cloud providers can be partially diluted. Even partial local deployment can change bargaining power by reducing dependence for baseline workloads.

Procurement and governance choices can also counteract consolidation. Large buyers can demand interoperability, transparent pricing, and flexible contracting. They can avoid long-dated commitments that eliminate options. They can invest in internal capability to evaluate and operate multiple model sources rather than outsourcing the entire stack to a single vendor.

Finally, regulation and public scrutiny can act as a counterforce, not by picking winners, but by limiting certain forms of bundling and by increasing disclosure about commitments and dependencies. Whether this becomes meaningful depends on enforcement and on how rapidly the technology and contracting norms evolve.

The key takeaway is that consolidation is not a story about model dominance alone. It is a story about surfaces, contracts, and cost structures. Google is a prominent candidate for a squeeze narrative because it controls multiple default surfaces and monetization rails. But the broader thesis is platform-general: in AI, incumbents can win by turning AI into the interface layer of existing products, while challengers must win by creating new workflows, new trust relationships, and credible portability.

8. Stress Tests and Simple Metrics

The AI boom invites story-driven thinking: new models, new demos, new proclamations of inevitability. If you want to evaluate crash risk or consolidation risk without getting trapped in narrative, you need a small set of stress tests and observable metrics that can be applied across firms and across layers of the stack. The goal here is not perfect forecasting. It is disciplined triage: identify which business models are robust under disappointment and which are fragile because they require multiple optimistic assumptions to hold simultaneously.

This section proposes four categories of simple tests. They are "simple" not because they are easy to compute from public data, but because they ask questions that remain stable even when specific technologies change.

8.1 Unit economics of inference and the price floor for compute

Inference is where AI becomes an ongoing business rather than a one-time build. Training is episodic; inference is operational. Most crash narratives become plausible when inference economics cannot support the fixed commitments that were made to scale.

A practical way to anchor inference economics is to think in three numbers:

- Revenue per unit of usage: what the customer pays per message, per query, per token, per task, or per seat (depending on the product).
- Cost per unit of usage: the compute cost plus overhead required to deliver that same unit reliably at target latency.
- Contribution margin: revenue minus variable costs.

These numbers are rarely disclosed cleanly, but you can still ask: does the business have a credible path to positive contribution margin at scale, or is it structurally subsidizing usage?

The key concept is the price floor for compute: the lowest sustainable price at which compute can be sold while still covering:

- electricity and cooling
- data-center overhead
- networking and storage
- depreciation of hardware (or lease payments)
- operational labor and reliability engineering
- a cost of capital appropriate to the risk

If the market price of inference falls below this floor, someone must absorb losses. Incumbents can sometimes absorb losses via cross-subsidy. Leveraged intermediaries generally cannot.

A simple stress test is therefore:

- Price compression test: assume inference prices fall by a large percentage over a short period (because of competition, commoditization, or bundling). Which business models remain viable?
- Cost-down test: assume hardware efficiency improves, but not fast enough to offset price declines. Who is left with margin compression?
- Mix-shift test: assume usage shifts toward lower-margin workloads (high context, high latency sensitivity, heavy retrieval, enterprise compliance). Does margin hold?

The most important distinction is between "AI as a product margin" and "AI as a retention feature." If a firm must earn margin directly from inference, it is exposed to price wars. If a firm can treat inference as a cost of keeping users, it can set prices in ways that squeeze competitors. This connects unit economics directly to consolidation risk.

8.2 Utilization sensitivity: what breaks when ramps miss

In a capital-intensive stack, utilization is everything. It is also the easiest place for reality to deviate from plans, because ramp schedules are forecasts, not guarantees.

A simple way to test utilization fragility is to treat commitments as fixed and demand as variable.

Ask four questions:

- How much capacity is fixed today? This includes reserved compute commitments, leased infrastructure, and power capacity charges.
- How fast does usage have to ramp for those commitments to be comfortably covered by operating cash flow?
- What happens if the ramp is delayed? Not "never happens," but delayed by quarters.
- What happens if usage grows but pricing falls, so revenue grows slower than usage?

You can formalize this as a ramp-miss stress test:

- Delay scenario: shift the expected ramp out by 6–12 months.
- Discount scenario: assume the market forces pricing down while usage rises.
- Both scenario: delay plus discount, which is common in competitive markets.

Then identify breakpoints:

- Liquidity breakpoint: when does cash run short?
- Covenant breakpoint: when do financing covenants become threatened?
- Renegotiation breakpoint: when does the firm need to renegotiate commitments to survive?
- Reputation breakpoint: when do partners or customers treat the firm as distressed, amplifying churn?

The core insight is that utilization risk is not linear. Once a business crosses a certain threshold—where fixed commitments become a large fraction of gross margin—small shocks can force big strategic changes: layoffs, capex freeze, asset sales, or emergency fundraising.

A second, often overlooked utilization metric is concentration-adjusted utilization. High utilization from one or two counterparties is not the same as high utilization across a diversified customer base. Concentration turns utilization into counterparty risk.

8.3 Credit signals: covenants, spreads, refinancing walls, concentration

Fast crashes are usually credit events. You do not need perfect information to see credit stress building. You need a small set of signals that reflect whether capital is becoming more expensive or more conditional.

Key credit signals include:

- Covenant tightness: are borrowers operating close to covenant limits? Tight covenants make downturns self-accelerating because they force action early.
- Refinancing walls: when do major maturities come due? A firm can look healthy today and still be fragile if it must refinance large obligations soon.
- Cost of capital: are spreads widening for the relevant class of borrower? Rising spreads are a real-time indicator of market doubt.
- Collateral haircuts: are lenders demanding more equity per dollar of equipment? Rising haircuts imply lenders distrust resale values and cash-flow stability.
- Concentration exposure: how dependent is the borrower on a small set of customers, suppliers, or investors? Concentration can turn a business issue at one counterparty into a solvency issue for another.

A practical “credit hygiene” checklist is:

- Liquidity runway: months of cash and committed credit under conservative usage assumptions.
- Maturity ladder: a simple schedule of when obligations come due.
- Customer concentration: share of revenue or contracted payments from the top few customers.
- Contract quality: how much contracted revenue is take-or-pay versus usage-contingent, and how much can be renegotiated.

These are not exotic. They are standard credit analysis tools. The novelty is that they now matter in what many people still talk about as a software sector.

One meta-signal is whether companies stop talking about growth and start talking about “capital discipline,” “efficiency,” “balance-sheet strength,” and “strategic focus.” Those phrases can be honest and prudent, but they often appear when the cost of capital has changed.

8.4 Competition signals: distribution capture vs model commoditization

The consolidation thesis requires different signals than the crash thesis. The question is not whether the market is overbuilt, but whether value is concentrating at the platform layer.

Here the key distinction is between distribution capture and model commoditization.

Signals of distribution capture include:

- AI features becoming default in search, browsers, operating systems, and productivity suites.
- user behavior shifting toward embedded AI rather than standalone AI destinations.
- enterprise procurement shifting toward suite bundles rather than best-of-breed point solutions.
- increasing reliance on identity, billing, and governance rails controlled by a few incumbents.

Signals of model commoditization include:

- comparable model quality across multiple providers for common tasks.
- rapid erosion of price premiums for access to “the best model.”

- open-weight models becoming good enough for many enterprise and on-prem deployments.
- differentiation shifting from model quality to product integration, latency, reliability, and domain adaptation.

A simple competition stress test is:

- If base models become interchangeable, who still wins? The answer is usually the firm with the strongest distribution and the best system integration.
- If distribution surfaces are controlled by a few incumbents, what is the survivable niche for challengers? The answer is usually vertical specialization, workflow ownership, or a portability-first enterprise strategy.
- If bundling becomes the dominant go-to-market move, which firms can resist? Usually those with mission-critical domain products and high switching costs at the workflow level, not at the model level.

You can summarize the competitive state of a sector with one question:

Is AI being bought as a standalone product, or being consumed as a bundled feature of an existing platform?

If the answer trends toward “bundled feature,” consolidation risk rises even if crash risk falls. The system can remain financially stable while still squeezing out many independent players.

The point of these stress tests is to make the conversation falsifiable in practice. You may not be able to compute every number precisely, but you can still apply the structure: unit economics, utilization sensitivity, credit conditions, and distribution versus commoditization. If all four categories look healthy, the crash narrative weakens. If two or three flash red, the crash narrative strengthens. If unit economics and credit look manageable but distribution capture accelerates, then consolidation becomes the dominant outcome even in a non-crash world.

9. Governance and Policy Options

If the AI boom is increasingly an infrastructure boom, then the governance questions resemble those of infrastructure markets: transparency, market integrity, competition policy, and resilience. The goal is not to “stop AI.” The goal is to keep a capital-intensive buildout from becoming fragile through opacity, to reduce incentives for self-referential demand signaling, and to preserve meaningful competition when default surfaces and bundling can quietly concentrate power.

Policy here does not mean only government regulation. It also includes investor standards, lender underwriting norms, enterprise procurement discipline, and voluntary disclosure practices. The most effective guardrails often arise from a convergence: investors demand better disclosure, lenders standardize covenants, enterprises demand portability, and regulators focus on the narrow cases where market power and deception risk are highest.

9.1 Disclosure: standardizing commitments, utilization, and counterparty risk

The single most constructive intervention in a fast-growing infrastructure market is better disclosure. Better disclosure does not prevent booms; it reduces the chance that booms become systemic surprises.

Three disclosure categories matter most.

Commitments. Companies should report, in a consistent way, the size and nature of long-dated obligations tied to compute and infrastructure. The point is not to reveal proprietary details, but to let stakeholders understand fixed obligations relative to cash flow. Useful standard fields include:

- total committed spend over specified horizons (1 year, 3 years, 5 years)
- the share that is take-or-pay or minimum spend versus usage-contingent
- termination rights and renegotiation triggers (high-level, not vendor-specific)
- prepayments and credits that effectively shift cash flow timing

Utilization. Utilization is the economic bridge between capacity and profit. In a market where commitments are large, stakeholders need visibility into whether capacity is actually being used and whether usage is profitable. Standardized utilization disclosure might include:

- utilization rates for owned or leased compute fleets (by broad workload category)
- the fraction of capacity under contract versus actually consumed
- ramp assumptions used internally for planning (in ranges)
- the share of revenue tied to a small number of counterparties' utilization behavior

Counterparty risk. Circularity becomes dangerous mainly when counterparties are concentrated and interdependent. Standard disclosure could include:

- top-customer concentration for contracted payments (in bands)
- exposure to any single supplier, cloud provider, or data-center landlord (in bands)
- the degree of financing dependence on a small set of investors or lenders

- related-party adjacency risk (not necessarily formal related parties, but ecosystems)

These disclosures matter because they convert “vibes” into a comparable risk profile. They also reduce the incentive for superficial narrative signaling. If the market can see commitments and utilization side by side, it becomes harder to use commitments alone as proof of durable demand.

9.2 Market integrity: limiting misleading demand signaling

Market integrity interventions aim at a narrow target: prevent the appearance of independent demand when demand is materially financed or backstopped by the same ecosystem that benefits from the demand signal.

This is not about banning strategic partnerships. It is about preventing confusion in the market about what a contract or commitment actually represents.

Governance options include:

Disclosure of financing linkage. When a large commitment is announced, stakeholders should be able to understand whether it is supported by supplier financing, investor credits, or other structures that materially reduce the buyer’s independent exposure. If the buyer’s ability to meet commitments depends on continued funding from the same ecosystem, that dependence should be visible.

Clear separation of bookings, commitments, and cash flow. Firms often report growth in ways that emphasize contracted commitments. That is not wrong, but it becomes misleading if audiences interpret commitments as realized, profitable usage. Investors and analysts can push for consistent separation:

- contracted commitments (future)
- realized usage (present)
- realized cash flow and margin (economic reality)

Guardrails on aggressive forward guidance. In fast-moving sectors, guidance becomes a narrative instrument. One integrity move is to standardize the assumptions behind “capacity plans” and to discourage the presentation of best-case growth as base case. This can be done through investor relations norms and by analysts insisting on downside scenarios.

Underwriting discipline. Lenders and private-credit markets can improve integrity by adopting standardized stress tests: utilization drops, pricing compression, collateral markdowns, and refinancing scenarios. When underwriting becomes more consistent, fragile structures become harder to finance at scale.

The pragmatic goal is not moral purity. It is to reduce the probability that self-referential financing chains masquerade as broad market pull.

9.3 Competition: interoperability, bundling scrutiny, and procurement rules

If consolidation is a plausible endgame, then competition policy is not a side issue. It is central. But the most constructive competition interventions are often targeted at interoperability and procurement rather than at trying to select winners.

Interoperability. A key counterforce to platform squeeze is the ability to move workloads and data. Competition policy can support interoperability by encouraging:

- portable data formats and clear export paths for key AI artifacts (logs, evaluation sets, embeddings where appropriate, fine-tuned weights where legally possible)
- standardized interfaces for model invocation and tool use, reducing the cost of multi-vendor strategies
- transparency about performance and pricing in comparable terms, so buyers can switch based on outcomes

Bundling scrutiny. Bundling is not inherently anti-competitive; it often reduces buyer friction. The concern arises when bundling uses dominance in one market to foreclose competition in another by making alternatives economically infeasible. Scrutiny should focus on patterns like:

- tying AI features to dominant distribution surfaces in ways that penalize alternatives
- cross-subsidized pricing that is below sustainable cost for standalone competitors
- exclusive default settings that are hard to change or hidden in UI design
- contract terms that restrict multi-vendor deployments through punitive pricing

Procurement rules. Large enterprises and governments are powerful shapers of market structure. Procurement can be designed to preserve competitive space by:

- requiring multi-vendor evaluation for critical AI use cases
- avoiding long-dated commitments that eliminate competitive options unless there is a clear scarcity rationale
- mandating portability requirements in contracts, including exit provisions and data export
- separating “seat licenses” from “AI usage charges” to prevent hidden cost shifts

These policies do not guarantee a diverse market, but they can prevent consolidation from becoming purely a function of default surfaces and lock-in rather than of performance and trust.

9.4 Safety and resilience: auditability without freezing innovation

Safety in an infrastructure-heavy AI world is not only about model behavior. It is also about operational resilience: preventing brittle systems from causing cascading failures, and ensuring that AI deployments remain auditable and governable in the environments where they matter.

Auditability. Enterprises and regulators will increasingly demand the ability to answer simple questions:

- What model version produced this output?
- What data sources were used in retrieval?
- What guardrails were active?
- What was the confidence or uncertainty handling?
- Who approved deployment changes?

A practical safety framework emphasizes logging, versioning, and traceability without requiring disclosure of proprietary model internals. The goal is accountability, not surveillance.

Resilience standards. Because AI services can become embedded into critical workflows, resilience standards matter:

- fallback modes when AI services are degraded
- clear rate limits and failure handling
- incident reporting norms for major failures
- stress tests for workload surges and dependency failures

Secure-by-default deployment. Many real harms arise not from models being “evil,” but from insecure integration: data leakage, prompt injection, tool misuse, and privilege escalation. Governance can focus on:

- secure tool invocation patterns
- least-privilege access to enterprise systems
- separation of duties for high-impact actions
- red-teaming and adversarial testing as standard practice

Avoiding innovation freeze. The policy risk is overreaction: heavy-handed rules that treat all AI deployments as equally dangerous, discouraging beneficial use. A more coherent approach is tiering:

- low-risk applications with lightweight requirements
- medium-risk applications with auditability and governance requirements
- high-risk applications with strict controls, independent evaluation, and incident obligations

This tiered approach aligns safety with actual impact rather than with generalized fear, and it allows innovation to continue while increasing accountability where it is needed most.

The core message of this section is that the best governance responses are not primarily ideological. They are engineering-like: standardize disclosure, reduce misleading demand signaling, preserve competition through portability and procurement discipline, and build safety through auditability and resilience. These measures do not prevent either crashes or consolidation, but they reduce unnecessary fragility, improve market clarity, and keep optionality alive in a rapidly industrializing sector.

10. Conclusion

The AI boom of 2024–2026 is not merely a software cycle. It is an infrastructure cycle: chips, data centers, power, and long-dated commitments. Once you see it this way, the loudest debates become easier to parse. “Circular deals” are not primarily a scandal category. They are a coordination mechanism in a scarce, urgent market where a small number of platforms control crucial bottlenecks. The system can be both real and fragile at the same time. It can also avoid a crash and still produce a squeeze, concentrating value in incumbents that control default distribution and cloud gravity.

This paper has tried to replace a moralized debate with a mechanical one: identify the structures, identify the assumptions they require, and identify the conditions under which those assumptions break. That framing matters because the most plausible failure mode is not an AI winter of capability. It is a mismatch between fixed commitments and cash flows, amplified by credit conditions and by the opacity of what “demand” actually means.

10.1 The central claim: circularity is incentive-consistent but can be brittle

The central claim can be stated simply.

Circularity is incentive-consistent because it helps rational actors solve real problems: secure scarce compute, accelerate time-to-ship, lock in capacity, and integrate into dominant distribution surfaces. It converts uncertain future demand into present-day commitments, which in turn unlocks supply, construction, and financing. It also aligns with platform strategy: incumbents can finance growth that returns as spend on their own infrastructure and appears as momentum.

But circularity can be brittle because the same coupling that enables rapid scaling also transmits shocks. When commitments are large, leverage is layered, counterparties are concentrated, and refinancing is assumed, then small changes in the environment can propagate quickly. A modest ROI disappointment can become a capex freeze. A mild credit tightening can become forced deleveraging. A normal price war can become collateral repricing. None of these require fraud. They require only the sector's standard optimism to collide with the standard physics of infrastructure finance.

This brittleness is not guaranteed. It is conditional. The task is not to denounce the loop. The task is to measure how much the loop is substituting for independent demand and how much downside is amplified by debt, commitments, and concentration.

10.2 What would falsify the crash narrative

"Crash" talk is cheap; falsification is harder. Still, there are clear conditions that would weaken the crash narrative substantially.

Sustained, improving inference unit economics would be a major falsifier. If AI services can deliver consistent positive contribution margins across broad workloads, even as prices fall, then the infrastructure buildout has a credible cash-flow foundation. Price compression would become manageable rather than destabilizing.

High, diversified utilization would also falsify the crash story. If fleets remain well utilized by a wide set of customers and workloads, the system has real demand pull rather than ecosystem self-funding. Utilization that is stable even as funding conditions tighten is especially important.

A third falsifier is a smooth transition from scarcity pricing to commodity pricing without widespread distress. If compute becomes cheaper and more available while leveraged intermediaries and landlords remain solvent and continue expanding in a disciplined way, then the market has absorbed the capex cycle without a hard break.

A fourth falsifier is stable refinancing and conservative leverage. If the neocloud and infrastructure layers maintain strong liquidity, avoid near-term refinancing cliffs, and show durable contract performance through a tighter credit regime, then the “fast crash” pathway loses force.

Finally, if incumbents bundle AI aggressively yet still leave meaningful room for profitable specialists and multi-vendor enterprise deployments, then the consolidation thesis weakens. It would imply a more competitive equilibrium than the squeeze story suggests.

10.3 What would confirm it early enough to matter

Confirmation, for decision-making purposes, means early warning signals that appear before a cascade is unavoidable.

The most actionable confirmatory signals cluster in four areas.

First, utilization and pricing: visible discounting for capacity, shortening contract durations, and a shift in buyer behavior toward flexibility rather than long commitments. If this is paired with slower ramp schedules, the slow-crash pathway strengthens.

Second, credit tightening: worsening financing terms, increased collateral haircuts, covenant stress, and a visible reluctance to finance new capacity. Credit-market posture often changes before public narratives do.

Third, forced renegotiations and project pauses: if major expansion projects are delayed, downsized, or quietly paused, especially when explained as “optimization” rather than as demand-driven, the system may be entering a digestion phase. If renegotiations occur under pressure rather than as normal refresh cycles, fragility is rising.

Fourth, concentration stress: if a few counterparties become recognized as systemically important—where their ability to fund or utilize capacity determines the viability of multiple intermediaries—then circularity has crossed from ecosystem-building into systemic coupling.

The key is not to wait for bankruptcy headlines. By the time those occur, the repricing has already happened. Early confirmation is a shift in contract behavior, credit terms, and utilization realities.

10.4 Practical next steps for analysts, operators, and policymakers

Different stakeholders can act without requiring unanimity or heavy-handed interventions.

For analysts and investors:

- Track obligations, not just growth. Focus on long-dated commitments relative to realistic cash flow, and ask how those commitments behave under ramp delays and price compression.
- Separate commitments from profitable usage. Treat booked commitments as a leading indicator, not as proof, until utilization and margin follow.
- Watch the credit layer. Credit conditions often signal turning points earlier than equity narratives. Look for refinancing walls, covenant pressure, and collateral markdown behavior.
- Stress-test concentration. Model what happens if the top few counterparties reduce spend or renegotiate. Concentration converts ordinary uncertainty into systemic risk.

For operators and enterprise buyers:

- Contract for flexibility. If scarcity forces commitments, build in exit provisions, portability, and multi-vendor options where possible.
- Measure contribution margin per workload. Separate “cool demo usage” from the workloads that pay, and scale only those that can carry their own costs.
- Avoid single-point dependency by default. Even if one vendor is best today, keep credible alternatives alive through architecture choices and procurement discipline.
- Plan for failure modes. Build fallback workflows and auditability so that AI integration does not become a single point of operational failure.

For policymakers and regulators:

- Standardize disclosure for commitments and utilization. The best systemic-risk reduction is transparency about fixed obligations, realized usage, and counterparty concentration.
- Focus integrity rules on clarity, not on banning partnerships. Require clear disclosure when financing materially supports demand signals.
- Preserve competition through portability and procurement. Interoperability and credible switching reduce squeeze dynamics more effectively than trying to pick winners.

- Promote resilience standards. Encourage auditability, incident reporting, and secure integration practices in high-impact domains without imposing blanket restrictions that freeze beneficial use.

The overall conclusion is not that an AI crash is inevitable, nor that consolidation must happen. The conclusion is that the current buildout has the structure of an infrastructure cycle, and infrastructure cycles have recognizable physics. Circularity can be a rational accelerator. It can also be a brittle coupling. The difference is observable in unit economics, utilization, credit conditions, and distribution capture. Those are the levers that allow decision-makers to move from narrative to measurement—and, crucially, to act while options still exist.

References

R1 Reporting and analysis on circular AI deals and round-tripping

- Wong, M., & Warzel, C. (2025, October 30). Here's How the AI Crash Happens. The Atlantic. The Atlantic
- Karma, R. (2025, December 10). Something Ominous Is Happening in the AI Economy. The Atlantic. The Atlantic
- Forgash, E., & Ghosh, A. (2025, October 7). OpenAI, Nvidia Fuel $1 Trillion AI Market With Web of Circular Deals. Bloomberg. Bloomberg
- Financial Times. (2025, October 6). OpenAI's computing deals to top $1tn as demand for AI surges. Financial Times
- Financial Times. (2025, February 6). Big Tech companies to spend $300bn on AI in 2025, says Morgan Stanley. Financial Times
- Reuters. (2025, September 15). CoreWeave orders Nvidia chips worth $6.3 billion as AI demand surges. Reuters
- Sequoia Capital. (2024, June 20). AI's $600B Question. Sequoia Capital
- Goldman Sachs. (2025, October 22). AI: In a Bubble. Goldman Sachs
- Yale Insights. (2025, October 8). This Is How the AI Bubble Bursts. Yale Insights
- Smith, N. (2025, October 22). Should we worry about AI's circular deals? Noahpinion. Noahpinion
- Zitron, E. (2025, August 6). AI Is A Money Trap. wheresyoured.at
- JPMorgan Asset Management. (2025). Note on "circularity" concerns in AI dealmaking and demand signaling. JPMorgan

R2 Cloud capex cycles, data-center economics, and power constraints

- International Energy Agency. (2024). Energy and AI (data centres, networks, and electricity demand framing). IEA
- CBRE. (2025, September 8). North America Data Center Trends (power availability and supply constraints). CBRE
- McKinsey & Company. (2025). Data centers: How to build bigger, faster, and cheaper (AI-driven capacity buildout). McKinsey & Company

- McKinsey & Company. (2025). The cost of compute: Powering AI (power as the binding constraint). McKinsey & Company
- Reuters. (2024, August 20). Data center construction and capacity pipeline dynamics in the US. Reuters
- Uptime Institute. (2025). Reporting on data center constraints and availability risks. Uptime Institute
- U.S. Energy Information Administration. (2024–2025). Coverage of data center load growth and grid impacts. EIA
- National Renewable Energy Laboratory. (2024). Work on load growth / grid planning implications relevant to data centers. The Department of Energy's Energy.gov
- S&P Global Ratings. (2024–2025). Credit implications of data center expansion and power constraints. S&P Global
- SemiAnalysis. (2024–2025). Neocloud economics, cluster capacity, and GPU infrastructure analysis. SemiAnalysis+1

R3 Private credit, structured finance, and systemic risk transmission

- International Monetary Fund. (2024, April). Global Financial Stability Report: The Rise and Risks of Private Credit. IMF
- Bank for International Settlements. (2024). Annual Economic Report (macro-financial vulnerabilities framing relevant to leverage/liquidity channels). Bank for International Settlements
- Financial Stability Board. (2025, December 16). Global Monitoring Report on Nonbank Financial Intermediation (data gaps and vulnerabilities, including private credit). Financial Stability Board
- Federal Reserve. (2025). Financial Stability Report (monitoring leverage, funding risk, and nonbank channels). Federal Reserve

R4 Competition and platform economics in search and cloud AI

- U.S. Federal Trade Commission. (2025, January). Staff report on AI partnerships and investments (competitive implications of major platform tie-ups). Federal Trade Commission
- Reuters. (2024–2025). Reporting on Google search distribution/defaults, antitrust remedies, and competitive effects. Reuters

- Bloomberg. (2025). Reporting on Gemini integration and distribution surfaces (Chrome/Android/Search) and competitive consequences. Bloomberg
- UK Competition and Markets Authority. (2024). AI foundation models work (principles, switching costs, and interoperability concerns). Hindenburg Research
- Rochet, J.-C., & Tirole, J. (2002/2003). Platform Competition in Two-Sided Markets. TSE
- Farrell, J., & Klemperer, P. (2007). Coordination and Lock-In: Competition with Switching Costs and Network Effects. Econometrics Laboratory+1
- Parker, G. G., & Van Alstyne, M. (2005). Two-Sided Network Effects: A Theory of Information Product Design. INFORMS Pubs Online
- Evans, D. S., & Schmalensee, R. (2016). Matchmakers: The New Economics of Multisided Platforms. Harvard Business Review Store

R5 Accounting, disclosure, and contract-risk literature

- U.S. Securities and Exchange Commission. (2020, November 19). Final rule: Management's Discussion and Analysis, Selected Financial Data, and Supplementary Financial Information (Release No. 33-10890). SEC+1
- Cornell Law School (Legal Information Institute). (Current CFR). 17 CFR § 229.303 (Item 303) MD&A requirements. Legal Information Institute
- Deloitte. (2025, November). Revenue Recognition Roadmap (ASC 606 and related disclosure/judgment topics). DART
- Deloitte. (ASC 842). Lessee disclosure requirements (lease commitments, restrictions, cash flow uncertainty). DART
- IFRS Foundation / IASB. (IFRS 16). IFRS 16 Leases (standard overview and current issued-standards page). IFRS+1
- Financial Accounting Standards Board. (2024). GAAP Taxonomy Implementation Guide on Leases under Topic 842. FASB
- Financial Accounting Standards Board. (2025, September). Post-Implementation Review materials for Topic 842 (disclosure/process implications). storage.fasb.org

The author has published over 4,800 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.